DOCUMENT RESUME

ED 365 713 TM 020 907

AUTHOR Kim, JinGyu
TITLE Individual Differences in Computerized Adaptive Testing.
PUB DATE Nov 93
NOTE 22p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (New Orleans, LA, November 10-12, 1993).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; *Adaptive Testing; Affective Behavior; Age Differences; Bayesian Statistics; Cognitive Processes; *Computer Assisted Testing; Computer Literacy; Demography; Ethnic Groups; *Individual Differences; *Item Response Theory; Mathematics Anxiety; Maximum Likelihood Statistics; Psychological Characteristics; Sex Differences; Student Characteristics; Test Anxiety; Test Items; Test Wiseness
IDENTIFIERS Academic Self Concept; *Self Adapted Testing; Testlets

ABSTRACT

Research on the major computerized adaptive testing (CAT) strategies is reviewed, and some findings are reported that examine effects of examinee demographic and psychological characteristics on CAT strategies. In fixed branching strategies, all examinees respond to a common routing test, the score of which is used to assign examinees to a second-stage test. The currently popular statistically branched adaptive strategies are based on item-response theory, and include maximum likelihood strategy and Bayesian strategy. Two alternative strategies are the use of self-adapted testing and testlet strategies. Examinee characteristic variables are divided into: (1) demographic variables; (2) computer-use variables; (3) test-taking strategy variables; (4) cognitive characteristics; and (5) affective characteristics. Although research on the relationship between examinee psychological characteristics and CAT has been inconclusive, the basic findings are that examinees of different ethnic, gender, age, grade, ability, academic self-concept, test anxiety, computer anxiety, math anxiety, and computer experience groups are differentially affected by the adaptive testing strategies. Implications for research and practice are discussed. (Contains 67 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 365 713

INDIVIDUAL DIFFERENCES IN COMPUTERIZED ADAPTIVE TESTING

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JinGyu Kim
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

JinGyu Kim

The University of Alabama

P.O. Box 870231

Tuscaloosa, AL 35487-0231

Annual Meeting

Mid-South Educational Research Association

New Orleans, Louisiana

November 10-12, 1993

TM020987

BEST COPY AVAILABLE

INDIVIDUAL DIFFERENCES IN COMPUTERIZED ADAPTIVE TESTING

Computerized adaptive testing (CAT) is an efficient and viable alternative to paper-and-pencil testing. Recent item response theory research and advances in microcomputer technology have indicated CAT can adapt during the test administration according to student performance on each test item. In a CAT, test items are selected according to an algorithm that attempts to maximize the efficiency of a test by providing the maximum amount of information about an examinee's ability with the minimum number of items.

There are some areas that require further investigation in computerized adaptive testing. Current popular research issues are the reliability and validity of CATs, the ordering of items, the ability estimation procedures, and the context of item administration in the estimation of parameters (Wise & Plake, 1989). Research has paid much attention to the efficiency and precision of the examinee's ability estimation. However, the research on the effects of examinees' demographic and psychological characteristics on CAT has been largely neglected. In other words, how individual differences among examinees are systematically related to the adaptive testing strategies have not been extensively studied. In fact, there have been more studies on examinees' individual differences in conventional tests or computerized tests (CT) than in computerized adaptive tests. Recently, a few researchers have paid attention to the individual differences in CATs (Buhr & Legg, 1989; Legg & Buhr, 1992; Vispoel & Rocklin, 1993). Therefore, a comprehensive review on individual differences is needed to suggest some guidelines for further investigations of CATs. The purpose of this paper is to critically review the research on the major computerized adaptive testing strategies and to report the findings of some studies that examined the effects of examinee demographic and psychological characteristics on the computerized adaptive testing strategies.

Computerized Adaptive Testing Strategies

Adaptive testing is an alternative procedure which matches different sets of test items to different examinees' previous responses to items or abilities during the administration of a test. Generally, an operational adaptive test requires four components: item pool, item selection procedure, scoring procedure, and stopping rules. There are different approaches to selecting items from a pool, beginning points, scoring for individuals, termination for different individuals (Assessment Systems Corporation, 1989; Kingsbury & Zara, 1989; Reckase, 1989; Powell, 1991; Weiss, 1985). In this paper, the following three adaptive testing strategies are discussed; fixed-branching strategies, statistically branched strategies, alternative strategies.

Fixed-Branching Adaptive Strategies

The typical examples of the fixed-branching adaptive strategies are the two-stage tests, pyramidal test, and the stradaptive test. The two-stage tests (Angoff & Huddleston, 1958; Betz & Weiss, 1973, 1974; Kim & Plake, 1993; Lord, 1980; Weiss, 1974) are composed of a routing test and a measurement test. All examinees first respond to a common routing test. The routing test is typically a test of average difficulty. The score on that test is then used to assign each examinee to a second-stage measurement test. Responses to both tests are used to arrive at a final score. In a two-stage adaptive test, individuals who answer all or most of the items correctly on the short routing test receive a measurement test of higher difficulty; individuals who answer about half of the items correctly on the routing test receive a second-stage measurement test of average difficulty, and individuals who answer only a few items correctly on the routing test receive a second-stage measurement test of low difficulty.

The most popular example of multistage tests is the pyramidal test (Bayoff, Thomas, & Anderson, 1960; Larkin & Weiss, 1974; Lord, 1970; Weiss, 1974). The pyramidal test consists of a set of items prestructured by difficulty into a structure resembling a pyramid. At the top of the

pyramid is an item of average difficulty. At the next stage of the test are two items, one of which is slightly more difficult than the first item, with the other item slightly less difficult than that item. The branching rule used in a pyramidal adaptive test is that the slightly more difficult item is administered following a correct response to an item, and the slightly easier item is administered following an incorrect response to an item.

The best example of this type of strategy is the stradaptive test (Vale & Weiss, 1975a, 1975b, 1978; Weiss, 1973) in which several subtests are defined, each containing items at a specified difficulty level. In a stradaptive test, testing proceeds by administering an item in one stratum and then branching to a more difficult stratum if the item is answered correctly or to a less difficult stratum if it is answered incorrectly. Whenever an examinee branches to a stratum, the next previously unadministered item in that stratum is administered to the examinee.

The fixed-branching adaptive strategies are useful if adaptive tests are to be administered by paper-and-pencil or by simple testing machines. However, these approaches to adaptive testing have several limitations (Weiss, 1985). A primary limitation is that they generally use only item difficulty information in order to structure the item pool. The second problem is how to develop scoring methods that can be used when different items are answered by different examinees. The third problem is that the strategies are designed for fixed-length test administration with the exception of the stradaptive test.

Statistically-Branched Adaptive Strategies

The currently popular approaches for adaptive testing were developed during the late 1960s and early 1970s based on item response theory (IRT) methodology. IRT is a statistical theory consisting of a family of models that express the probability of observing a particular response to an item as a function of certain characteristics of the item and of the ability level of the examinee (Crocker & Algina, 1986; Hambleton & Swaminathan, 1985). The typical IRT-

based strategies are maximum likelihood strategy and Bayesian strategy (Weiss & Kingsbury, 1984).

Maximum Likelihood Strategy: The likelihood function for a set of test items indicates the probability of observing the entire vector of obtained item responses at each level of ability. From this likelihood function, an estimate of the examinee's ability can be obtained. Conceptually, this can be done by assuming that the best estimate of an examinee's ability is the level of ability that would most likely produce the vector of responses observed. This is determined by locating the maximum value of the likelihood function and identifying the ability level (theta) associated with that maximum. This score is called the maximum likelihood estimate of ability. The maximum information adaptive testing strategy (Weiss, 1982, 1985) selects items that provide maximum levels of item information at an individual's currently estimated trait level. After the administration of an item and estimation of trait level, the new trait level is used to select the next item to be administered to that examinee. In a maximum information adaptive test, a sequential process is specified in which an item is administered, an ability estimate is calculated, the item providing the most information at that estimate is selected, and the process is repeated. This process will continue until a fixed number of items has been administered or until some other criterion for termination has been satisfied.

Bayesian Strategy: A Bayesian estimate (e.g., Owen, 1969, 1975) is conceptually very similar to the maximum likelihood estimate. Bayesian strategies use Bayes' theorem to estimate an examinee's ability. Bayes theorem generates a posterior probability distribution from the combination of a prior probability distribution and the current observation. The Bayesian posterior likelihood function can become a legitimate probability density function with a mean and variance. The mean or the mode of the posterior can be taken as an estimate of ability. In Bayesian strategies, items are selected on the basis of minimizing the Bayesian posterior variance

of the ability estimate rather than maximizing values of item information (Owen, 1969, 1975). Owen's item selection strategy utilizes a current ability estimate and its Bayesian variance as the prior distribution for the item selection process. In order to select the next item to be administered during the adaptive test, this method evaluates the posterior variance of the ability estimate for each item in the pool under two conditions: a) if the item is answered correctly and b) if the item is answered incorrectly (Weiss, 1982).

According to Weiss and Kingsbury (1984), two efficient item selection procedures are maximum information and Bayesian. Both procedures involve searching the entire pool of unadministered items for a single item. Because of the relationships between information and Bayesian posterior variance, maximum information strategy and Bayesian strategy will frequently select a similar subset of items in many cases (Sympson, Weiss, & Ree, 1982). In obtaining ability estimates, maximum likelihood estimation poses problems when the number of test items is small. Bayesian procedures overcome the problems encountered with maximum likelihood procedures but may produce biased estimates of ability if inappropriate prior distributions are chosen (Hambleton, Swaminathan, & Rogers, 1991).

Alternative Adaptive Strategies

In this section, two alternatives to the traditional adaptive strategies that merit discussion are the use of self-adapted testing and testlet strategy (Kingsbury & Zara, 1989; Plake, 1993).

Testlet Strategy: The concept of the testlet was introduced explicitly by Wainer and Kiely (1987) as content-based item clusters which are analyzed as units and are independent of all other testlets and items. Thus, the testlet was proposed as the unit of construction and analysis for CATs. It could ease some of the observed and prospective difficulties associated with most current methods of test construction such as context effects, item ordering, and content balancing (Wainer, et al., 1990).

The procedures of the testlet strategy are as follows. First, specify an initial estimate of proficiency (this specifies an initial testlet). Second, estimate proficiency after each testlet. Choose the remaining testlet that is most informative near the estimated proficiency to be administered next. Finally, stop when the precision of the estimated proficiency is adequate, or when some pre-specified number of testlets have been administered.

Testlets have been applied to scaling of reading comprehension items (Thissen et al., 1989), to the measurement of algebra items (Wainer & Lewis, 1990; Wainer, Lewis, Kaplan, & Braswell, 1991), and to scaling performance assessment tasks (Yen, 1992). Although applications are possible, it is unlikely that a testlet-based adaptive test could be used in most ongoing testing situations (Kingsbury & Zara, 1989).

Self-Adapted Testing: Rocklin and O'Donnell (1987) developed an alternative procedure, called self-adapted testing, in which examinees could choose the difficulty of the items they attempt on an item-by-item basis. This strategy seeks to minimize student anxiety and maximize student performance by allowing the examinee to choose items, rather than by a computer algorithm (Rocklin & O'Donnell, 1991; Wise, Plake, Johnson, & Roos, 1992; Wise, Roos, Plake, & Nebelsick-Gullett, 1993).

In self-adapted testing, the student takes a test question, is informed whether his or her answer was correct, and then decides how difficult the next item on test should be. To facilitate student decisions concerning item difficulty, the items are prestructured into difficulty groups or strata, as in the stradaptive test. The major difference between the stradaptive test and self-adapted test is that the examinee chooses the difficulty stratum, rather than having the computer choose it (Kingsbury & Zara, 1989).

Some research findings have shown that examinees taking the self-adapted test scored significantly higher than those taking the computerized adaptive test (Rocklin & O'Donnell, 1987;

Vispoel & Rocklin, 1993; Wise, Plake, Johnson, & Roos, 1992). Although this strategy might reduce test anxiety, it would produce a test score with a very low information value because the student has option to take a test is far below (or above) his or her optimal performance level (Kingsbury & Zara, 1989).

Examinee Characteristics Variables

There is some concern that computerized tests, including CAT, may produce differential effects for different groups of students (Legg & Buhr, 1992). There are several examinee demographic and psychological characteristics that can contribute to computerized tests. Parshall and Kromrey (1993) have divided examinee characteristics into the following three sub-categories; (a) demographic variables (gender, racial/ethnic background, and age), (b) computer use variables (variety of computer experience, frequency of computer use, frequency of mouse use, and test mode preference), and (c) test taking strategy variables (test strategy preference, tendency to omit items, and tendency to review item). They did not include examinee cognitive and affective characteristics in their classification.

In this paper, examinee characteristics are divided into five sub-categories by adding these two characteristics. Each specific variables indicating individual difference can be found in typical computerized testing studies. First, demographic variables include gender, race or ethnic background, grade, and age (Johnson & Mihal, 1973; Johnson & White, 1980; Sorensen, 1985; Llabre & Froman, 1987; Moe & Johnson, 1988; Parshall & Kromrey, 1993). Second, cognitive characteristics variables belong to ability, aptitude, and achievement (Lee, Moreno, & Sympson, 1986; Wise & Wise, 1987). Third, affective characteristics variables contain anxiety, test anxiety, computer anxiety, math anxiety, self-concept, and attitudes (Wise, Plake, Eastman, Boettcher, & Lukin, 1986; Moe & Johnson, 1988; Wise, Barnes, Harvey, & Plake, 1989). Fourth, computer use variables include computer experience, computer use, and mouse use (Lee, 1986; Wise, Barnes,

Harvey, & Plake, 1989; Dimock & Cormier, 1991; Parshall & Kromrey, 1993). Finally, the examples of test taking strategy variables are test strategy preference, tendency to omit items, tendency to review items, response time, and testwiseness (Rocklin & O'Donnell, 1987; Spray, Ackerman, Reckase, & Carlson, 1989; Ward, Hooper, & Hannafin, 1989; Wise & Plake, 1989; Green, 1991).

However, a few studies (Rocklin & O'Donnell, 1987; Legg & Buhr, 1992) have begun to investigate the relationship between examinee characteristics and CATs. It is assumed that two approaches search for individual differences in terms of the testing situation. The first approach is trying to investigate examinee differences in the typical computerized adaptive testing situation. The second approach explores individual differences in the self-adapted testing, each examinee is allowed to choose the level of difficulty of the next item to be presented from among several levels of difficulty. In this paper, examinees' individual differences will be described by the above mentioned variables both in the CAT and in the self-adapted testing.

Demographic Variables

The examinee's demographic variables include gender, age, ethnic background, and grade. Olsen, Maynes, Slawson, and Ho (1989) compared the effectiveness of paper-administrated, computer-administrated, and computerized adaptive achievement tests for grades three and six. This was a pioneer study to evaluate computerized adaptive testing at the elementary grade school level. The investigators found no significant differences between paper-administered tests and computer-administered tests for grades three and six.

Legg and Buhr (1992) looked at how administration of the examination by computer affected examinees and, in particular, whether examinees of different ethnic, gender, age, ability, and computer-experience groups were differentially affected. They also explored whether group differences in reactions to the computerized test administration could help to explain observed

differences between CAT and conventional test scores and differences in the time used for testing. They developed three adaptive tests which consisted of mathematics, reading, and writing, using the MicroCAT (Assessment System Corporation, 1989) software program. The study used the questionnaire which consisted of 19 Likert-type items and 4 open-ended questions. The investigators found that little difference was observed between examinees less than 30 years of age and those 30 years old and older in their response to the questionnaire. They also found that mean scores for males and females differed in two areas. First, male students responded significantly less favorably than females. Second, males were more critical than females of the graphics for the mathematics items. They showed that mean scores for White, Black, Hispanic and Asian ethnic groups differed for several questions. First, Hispanic and Asian students were much less likely to indicate that they had had enough practice in responding than White students. Second, Asian students reported greater eye strain at the end of the test than other ethnic groups. Third, Asian students also differed from other students in their preference for the computer test and preferred the regular test, in contrast to the other groups.

Cognitive Characteristics Variables

Cognitive characteristic variables include ability, aptitude, and achievement. In general the related literature is fairly sparse because examinees' abilities, aptitudes, and achievements are largely used as dependent variables in the CAT research. Schinoff and Steed (1988) investigated that lower proficiency examinees liked a CAT better because they did not feel the discouragement associated with facing a long string of items that were too hard for them.

Legg and Buhr (1992) divided examinee into three ability groups based on conventional reading test scores and defined "low ability" as more than one standard deviation below the mean, "high ability" as more than one standard deviation above the mean, and "average" as within one standard deviation of the mean. They found that higher-ability examinees were more bothered

than lower-ability examinees by not being able to review items after completing them.

Affective Characteristics Variables

Examinees' affective characteristics which have been investigated in CAT research are test anxiety (Powell, 1991; Vispoel & Rocklin, 1993), computer anxiety (Legg & Buhr, 1992; Vispoel & Rocklin, 1993), math anxiety (Wise, Roos, Plake, & Nebelsick-Gullett, 1993), self-concept (Vispoel & Rocklin, 1993), and attitudes toward computerized adaptive tests (Schmidt, Urry, & Gugel, 1978; Moe & Johnson, 1988; Vispoel & Rocklin, 1993). Schmidt, Urry, and Gugel (1978) investigated examinee reactions to computer assisted tailored testing. They reported the reactions and opinions of 163 examinees who participated in a tailored pilot study conducted at the U.S. Civil Service Commission during the fall of 1975. They showed that the reactions of the examinees were positive.

Moe and Johnson (1988) studied participants' reactions to computerized adaptive ability test and assessed the practicability of this testing method in the classroom. Three hundred, fifteen students took a computerized and printed version of a standardized aptitude test battery, and a survey assessing their reactions. They found that overall reactions to the computerized test were overwhelmingly positive.

Powell (1991) examined the relationship between test anxiety and test performance in the three computerized adaptive testing procedures. She found no statistically significant differences among mean student achievement scores, nor among in-test anxiety means under the three adaptive testing methods. The study also showed that students who reported higher pre-test anxiety scored significantly higher in the matched-selection test, and students who preferred the matched selection and self-selection tests were significantly less anxious during those tests.

Legg and Buhr (1992) found a significant interaction between ethnic and gender groups on computer anxiety. In other words, differences in reported computer anxiety for the ethnic

groups were much larger for females than males, with Black females reporting the greatest anxiety and White females the least. For males, the mean for Hispanics was highest, while the mean for Asian examinees was lowest. In contrast, Hispanic females reported low computer anxiety. The mean indicated computer anxiety did not differ greatly for White and Black males. Contrary to expectations, females as a group reported less computer anxiety than males.

Vispoel and Rocklin (1993) assessed the effects of several individual difference variables (test anxiety, verbal self-concept, computer usage, and computer anxiety) on ability estimates alone and in interaction with the test administration procedures (maximum information adaptive, self-adapted, and fixed-item computerized vocabulary tests). The study used the same large, well-calibrated item bank for all the tests. They found significant main effects for test anxiety and self-concept variables and a significant self-concept by testing condition interaction. The most striking differences among administration conditions occurred for individuals with low verbal self-concepts, who performed noticeably better on the self-adapted test than on the other tests. As expected, estimated ability scores across the total sample were significantly higher for individuals with higher verbal self-concept and lower test anxiety.

Wise, Roos, Plake, and Nebelsick-Gullett (1993) investigated the relative influences of test type and test choice on examinee anxiety. They found that examinees low in math anxiety showed a strong preference for CAT, while the highly math anxious examinees chose self-adapted tests.

Computer Use Variables

Several studies have investigated the relationship between examinee performance on computerized tests and computer use variables (Lee, 1986; Wise, Barnes, Harvey, & Plake, 1989; Dimock & Cormier, 1991). Legg and Buhr (1992) also investigated computer experience by the self-report methods in the CAT testing situation. They found that computer experience did not differ significantly for gender or age groups, while differences were reported for ethnic and ability

groups.

Vispoel and Rocklin (1993) also assessed computer usage by participants' estimates of average number of hours a week they spend working on computers. They transformed the scores for computer usage responses by adding one to each score and taking the base 10 logarithm of the sum because their distribution was positively skewed. The study found no significant main effect for computer usage variable.

Test Taking Strategy Variables

The examinees' test taking variables include test taking strategy (Rocklin & O'Donnell, 1987; Wise, Plake, Johnson, & Roos, 1992; Wise, Roos, Plake, & Nebelsick-Gullett, 1993), response time (Gershon, Bergstrom, & Lunz, 1993), and item review (Lunz, Bergstrom, & Wright, 1992; Lunz & Stahl, 1993). Rocklin and O'Donnell (1987) explored a variant application of IRT in computerized testing, termed self-adapted testing, in which the difficulty levels of the items administered are chosen by the examinee, rather than by a computer algorithm. They found that examinees who received a self-adapted test scored significantly higher than examinees receiving a conventional computerized test.

Subsequent research studies (Rocklin & O'Donnell, 1991; Wise, Plake, Johnson, & Roos, 1992; Roos, Plake, & Wise, 1992) have explicitly compared self-adapted test and CAT. The studies described above indicated that a self-adapted test had yielded higher mean examinee test performance than a CAT, and had been accompanied by lower mean post-test state anxiety.

Gershon, Bergstrom, and Lunz (1993) analyzed the total response time for each item in CAT. They also divided total response time into initial test time and review time. They found that response time increased proportionately with increasing item text length and increasing item difficulty. The study showed that item sequence also was an important factor in that response time was greater for earlier items in the test.

FINDINGS AND IMPLICATIONS

Although the previous research on the relationship between examinee psychological characteristics and CAT has been inconclusive, the basic findings are that examinees of different ethnic, gender, age, grade, ability, academic self-concept, test anxiety, computer anxiety, math anxiety, and computer experience groups were differentially affected by the adaptive testing strategies in the related studies. In this section, the findings will be discussed in terms of five sub-categories.

First, research on the relationship between various demographic variables and CAT has not been conclusive. Gender, ethnic background, age, and grade are among the demographic variables which have been investigated. Some differences were observed between ethnic, gender, and age groups in their reactions to CAT, but these differences did not appear to affect the examinees' performance on the test. Although decisive evidence of the relationship between these variables and CAT has not been obtained, grounds for concern can be found in the results of national surveys on the equity of computer access (Becker & Sterling, 1987; Martinez & Mead, 1988). The lower access to the computers could cause an impact on the performance in CAT.

Second, cognitive characteristics variables have been largely used as dependent variables in the CAT research. One of the findings on these variables is that lower ability examinees did not indicate any greater problems in CAT than higher ability examinees. Based on the studies which found the difference of ability groups in their reactions to CAT, we can say that lower ability examinees have positive attitudes toward CAT.

Third, affective characteristics variables have been used as both dependent variables and independent variables in the CAT research. The findings were that (a) examinee attitude toward CAT was generally very positive, (b) computer anxiety in the CAT testing situation was related to examinees' ethnic and gender group, (c) test anxiety and verbal self-concept were significantly

related to ability estimates, with higher scores obtained by individuals with higher verbal self-concepts and lower test anxiety, and (d) a strong relationship was found between examinee test type choice and math anxiety level.

Fourth, some aspects of computer use variables have been investigated as potential sources of CAT, but the clear-cut results have not been obtained. A pattern of lower scores for examinees with less computer experience is frequently seen, although the score differences are often not statistically significant (Wise, Barnes, Harvey, & Plake, 1989). In other words, those students with less computer experience did not feel that they had enough practice responding, in comparison to the other two groups (Legg & Buhr, 1992).

Finally, test-taking strategy variables have been investigated as potential sources of CAT. The findings were that a self-adapted test had yielded higher mean scores than a CAT, and had been accompanied by lower mean post-test state anxiety. Specifically, the interaction of examinees' test-taking strategies with test flexibility appears to be important.

The research findings on examinee characteristics related to computerized adaptive testing strategies have some implications for test specialists and researchers. Although examinee characteristics may partially account for test performance in the CAT, some researchers also think that examinee characteristics play an important role in exploring the practicability of CAT in the near future.

First, the research findings have implications for equity issues in testing. The most extensive research has been conducted on the equivalence between computer-based tests and their conventional test counterparts. There is a paucity of research on the equivalence between CAT and CT or conventional test counterparts. These research may identify individual difference variables that influence the equivalence among the three modes of testing.

Second, the research on examinee characteristics provides the empirical evidences of test

validity. Validity refers to the extent to which an interference made from test scores is appropriate or meaningful (AERA/APA/NCME, 1985). Only by examining the relationships between scores on a test and the other variables specified in the theory can the validity of test administration procedures be compared (Vispoel & Rocklin, 1993). The relationship between examinee characteristics and CAT strategies provides some minimal information about the construct-related validity of the test administration procedures.

Third, the findings suggest that more practice items add to the actual tests for some subgroups of examinees. Adding one or two practice items would give examinees more opportunity to master the scrolling and might alleviate this problem.

Finally, cultural differences in CAT should be explored through the cross-cultural studies. There are no cross-cultural studies on examinees' performance and attitudes toward CAT. These studies will enable us to glimpse some possibilities that the CAT can be widly used from all over the world. This issue should be paid primary attention to future investigations of computerized adaptive testing.

REFERENCES

AERA/APA/NCME (1985). Standards for educational and psychological testing. Washington, DC: APA.

Angoff, W.H., & Huddleston, E.M. (1958). The multi-level experiment: A study of a two-level test system for the College Board Scholastic Aptitude Test (Statistical Report No. SR-58-21). Princeton, NJ: Educational Testing Service.

Assessment Systems Corporation (1989). User's manual for the MicroCAT testing system, version 3. St. Paul, MN: Author.

Bayroff, A.G., Thomas, J.J., & Anderson, A.A. (1960, January). Construction of an experimental sequential item test (Research Memorendum 60-1). Washington, DC: Personnel Research Branch, Department of Army.

Becker, H.J., & Sterling, C.W. (1987). Equity in school computer use: National data and neglected considerations. Journal of Educational Computing Research, 3, 289-311.

Betz, N.E., & Weiss, D.J. (1974). Simulation studies of two-stage ability testing (Research Report 74-4). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Betz, N.E., & Weiss, D.J. (1973). An empirical study of computer administered two-stage ability testing (Research Report 73-4). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Buhr, D.C., & Legg, S.M. (1989). Development of an adaptive test version of the College Level Academic Skills Test. (Institute for Student Assessment and Evaluation, Contract No. 88012704). Gainesville, FL: University of Florida.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York, NY: Holt, Rinehart and Winston.

Dimock, P.H., & Cormier, P. (1991). The effects of format differences and computer experience on performance and anxiety on a computer-administered test. Measurement and evaluation in Counseling and Development, 24(3), 119-126.

Gershon, R.C., Bergstrom, B.A., & Lunz, M. (1993, April). Computer adaptive testing: Exploring examinee response time. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Green, B.F. (1991). Guidelines for computer testing. In T.B. Gutkin & S.L. Wise (Eds.), The computer and the decision-making process(pp 245-254). Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage Publications.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff.

Johnson, D.F., & Mihal, W.L. (1973). Performance of blacks and whites in computerized versus manual testing environments. American Psychologist, 28, 694-699.

Johnson, D.F., & White, C.B. (1980). Effects of training of computerized test performance in the elderly. Journal of Applied Psychology, 65, 357-358.

Kim, H., & Plake, B.S. (1993, April). Monte carlo simulation comparison of two-stage testing and computerized adaptive testing. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Kingsbury, G.G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education, 2(4), 359-375.

Larkin, K.C., & Weiss, D.J. (1974, July). An empirical investigation of computer-administered pyramidal ability testing (Research Report 74-3). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Lee, J.A. (1986). The effects of past computer experience on computerized aptitude test performance. Educational and Psychological Measurement, 46, 727-733.

Lee, J.A., Moreno, K.E., & Sympson, J.B. (1986). The effects of mode of test administration on test performance. Educational and Psychological Measurement, 46, 467-474.

Legg, S.M., & Buhr, D.C. (1992). Computerized adaptive testing with different groups. Educational Measurement: Issues and Practice, 11(2), 23-27.

Llabre, M.M., & Froman, T.W. (1987). Allocation of time to test items: A study of ethnic differences. Journal of Experimental Education, 55, 137-140.

Lord, F.M. (1980). Applications item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F.M. (1970). Some test theory for tailored testing. In W. Holzman (Ed.), Computer-assisted instruction, testing, and guidance, New York, NY: Harper and Row.

Lunz, M.E., & Stahl, J.A. (1993, April). Test targeting and precision before and after review on computerized adaptive tests. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Lunz, M.E., Bergstrom, B.A., & Wright, B.D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. Applied Psychological Measurement, 16(1), 33-40.

Martinez, M.E., & Mead, N.A. (1988). Computer competence: The first national assessment. Princeton, NJ: Educational Testing Service.

Moe, K.C., & Johnson, M.F. (1988). Participants' reactions to computerized testing. Journal of Educational Computing Research, 4(1), 79-86.

Olsen, J.B., Maynes, D.D., Slawson, D., & Ho, K. (1989). Comparisons of paper-administered, computer-administered and computerized adaptive achievement tests. Journal of Educational Computing Research, 5(3), 311-326.

Owen, R.J. (1975). A Bayesian sequential procedure for quintal response in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.

Owen, R.J. (1969). A Bayesian approach to tailored testing (Research Report No. RR-69-92). Princeton, NJ: Educational Testing Service.

Parshall, C.G., & Kromrey, J.D. (1993, April). Computer testing versus paper-and-pencil testing: An analysis of examinee characteristics associated with mode effect. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Plake, B.S. (1993). Applications of educational measurement: Is optimum optimal? Educational Measurement: Issues and Practice, 12(1), 5-10.

Powell, Z-H. E. (1991). Test anxiety and test performance under computerized adaptive testing methods, Doctoral dissertation, Indiana University Graduate School, Bloomington, Indiana.

Reckase, M.D. (1989). Adaptive testing: The evolution of a good idea. Educational Measurement: Issues and Practice, 8(3), 11-15.

Rocklin, T., & O'Donnell, A.M. (1991, April). An empirical comparison of self adapted and maximum information item selection. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Rocklin, T., & O'Donnell, A.M. (1987). Self-adaptive testing: A performance-improving variant of computerized adaptive testing. Journal of Educational Psychology, 79(3), 315-319.

Roos, L.L., Plake, B.S., & Wise, S.L. (1992, April). The effects of feedback in computerized adaptive and self-adapted tests. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Schinoff, R.B., & Steed, L. (1988). The CAT program at Miami-Dade Community College. In D. Doucette (Ed.), Computerized adaptive testing: The state of the art in assessment at three community colleges (pp. 25-36). Laguna Hills, CA: League for Innovation in the Community College.

Schmidt, F.L., Urry, V.W., & Gugel, J.F. (1978). Computer assisted tailored testing: Examinee reactions and evaluations. Educational and Psychological Measurement, 38(2), 265-273.

Sorensen, H.B. (1985). Cognitive ability tests. AEDS Monitor, 24, 22-26.

Spray, J.A., Ackerman, T.A., Reckase, M.D., & Carlson, J.E. (1989). Effect of the medium of

item presentation on examinee performance and item characteristics. Journal of Educational Measurement, 26(3), 261-271.

Sympson, J.B., Weiss, D.J., & Ree, M.J. (1984, April). Predictive validity of computerized adaptive testing in a military training environment. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Thissen, D., Steinberg, L., & Mooney, J.A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. Journal of Educational Measurement, 26(3), 247-260.

Vale, C.D., & Weiss, D.J. (1978). The stratified adaptive ability test as a tool for personnel selection and placement. TIMS Studies in the Management Sciences, 8, 135-151.

Vale, C.D., & Weiss, D.J. (1975a, October). A study of computer-administered stradaptive ability testing (Research Report 75-4). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Vale, C.D., & Weiss, D.J. (1975b, December). A simulation study of stradaptive ability testing (Research Report 75-6). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Vispoel, W.P., & Rocklin, T. (1993, April). Individual differences and test administration procedures: A comparison of fixed-item, adaptive, and self-adapted testing. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. Educational Measurement: Issues and Practice, 12(1), 15-20.

Wainer, H., & Kiely, G.L. (1987). Item cluster and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24(3), 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. Journal of Educational Measurement, 27(1), 1-14.

Wainer, H., Lewis, C., Kaplan, B., & Braswell, J. (1991). Building algebra testlets: A comparison of hierarchical and linear structures. Journal of Educational Measurement, 28(4), 311-323.

Wainer, H., Doran, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Ward, T.J., Jr., Hooper, S.R., & Hannafin, K.M. (1989). The effects of computerized tests on the performance and attitudes of college students. Journal of Educational Computing Research, 5, 327-333.

Weiss, D.J. (1985). Adaptive testing by computer. Journal of Consulting and Clinical Psychology, 53(6), 774-789.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.

Weiss, D.J. (1974, December). Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D.J. (1973, September). The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D.J., & Kingsbury, G.G. (1984). Bias and information of Bayesian adaptive testing. Applied Psychological Measurement, 8, 273-285.

Wise, S.L., & Plake, B.S. (1989). Research on the effects of administering tests via computers. Educational Measurement: Issues and Practice, 8(3), 5-10.

Wise, S.L., & Wise, L.A. (1987). Comparison of computer-administered and paper-administered achievement tests with elementary school children. Computers in Human Behavior, 3, 15-20.

Wise, S.L., Barnes, L.B., Harvey, A.L., & Plake, B.S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. Applied Measurement in Education, 2(3), 235-241.

Wise, S.L., Plake, B.S., Eastman, L.A., Boettcher, L.L., & Lukin, M.E. (1986). The effects of item feedback and examinee control on test performance and anxiety in a computer-administered test. Computers in Human Behavior, 2, 21-29.

Wise, S.L., Plake, B.S., Johnson, P.L., & Roos, L.L. (1992). A comparison of self-adapted and computerized adaptive tests. Journal of Educational Measurement, 29(4), 329-339.

Wise, S.L., Roos, L.L., Plake, B.S., & Nebelsick-Gullett, L.J. (1993, April). The role of anxiety in examinee performance for self-adapted testing. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Yen, W.M. (1992). Scaling performance assessments: Strategies for managing local item dependence. NCME Invited Presentation in 1992 NCME Meeting.